

DOCUMENT RESUME

ED 331 876

TM 016 425

AUTHOR Clauser, Brian E.; And Others
TITLE Examination of Various Influences on the Mantel-Haenszel Statistic.
PUB DATE Apr 91
NOTE 17p.; Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, IL, April 3-7, 1991).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Comparative Testing; *Difficulty Level; *Item Bias; Item Response Theory; *Simulation; *Test Construction; Test Items
IDENTIFIERS *Mantel Haenszel Procedure

ABSTRACT

Item bias has been a major concern for test developers during recent years. The Mantel-Haenszel statistic has been among the preferred methods for identifying biased items. The statistic's performance in identifying uniform bias in simulated data modeled by producing various levels of difference in the (item difficulty) b-parameter for reference and focal groups was studied. Data were simulated for a 75-item test (59 non-biased and 16 biased items) with five administrations and 1,000 examinees in each group. Levels of difference in difficulty were crossed with four levels of the (item discrimination) a-parameter, five levels of the b-parameter, and two ability distributions. The results show significant effects for the difference in difficulty levels, and for the a-parameter and b-parameter values. Data suggest that: the Mantel-Haenszel statistic may differentially favor identification of more highly discriminating items, and the procedure is likely to be unable to detect bias in very difficult items. Recommendations for practitioners are provided. Four tables and four graphs present the simulated data. (Author/SLD)

* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *

Examination of Various Influences on the Mantel-Haenszel Statistic

Brian Clauser, Kathleen M. Mazor, Ronald K. Hambleton
University of Massachusetts at Amherst

Abstract

Item bias has been a major concern for test developers during recent years. The Mantel-Haenszel statistic has been among the preferred methods for identifying biased items. The present research examines the statistic's performance in identifying "uniform bias" in simulated data modeled by producing various levels of difference in the (item difficulty) b-parameter for reference and focal groups. These levels of difference in difficulty are crossed with four levels of the (item discrimination) a-parameter, five levels of the b-parameter, and two ability distributions. The results show significant effects for each of these variables. The data suggest that the Mantel-Haenszel statistic may differentially favor identification of more highly discriminating items. It also suggests that the procedure is likely to be unable to detect bias in very difficult items. Recommendations for practitioners are made.

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)

☐ This document has been reproduced as
received from the person or organization
originating it.

☒ Minor changes have been made to improve
reproduction quality.

• Points of view or opinions stated in this docu-
ment do not necessarily represent official
OERI position or policy.

"PERMISSION TO REPRODUCE THIS
MATERIAL HAS BEEN GRANTED BY

RON HAMBLETON

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)."

BEST COPY AVAILABLE

Examination of Various Influences on the Mantel-Haenszel Statistic^{1,2}

Brian Clauser, Kathleen M. Mazor, Ronald K. Hambleton
University of Massachusetts at Amherst

Bias in items used in standardized testing remains a very significant concern for test developers in spite of more than a decade of intense effort to understand and control this phenomenon. Researchers have used both judgmental and statistical approaches to examine the differential functioning of test items. A significant indicator of the inadequacy of the current state of the art for these approaches is the common failure to demonstrate some degree of convergent validity for the two techniques. Although this may be due to the insensitivity of the judgmental approaches, the question of exactly how the statistical procedures are functioning remains a major concern.

The present study attempts to add to the current understanding of how the Mantel-Haenszel (MH) statistic performs in the identification of differentially functioning items (Holland & Thayer, 1988). This statistic is among the preferred approaches for such item identification because it is relatively simple and inexpensive to perform and produces results that are similar to the more theoretically complex and expensive IRT approaches. Numerous studies have examined the Mantel-Haenszel statistic by comparing its performance with other bias detection procedures (Hambleton & Rogers, 1989) or by manipulating variables of interest such as the number of score groups (Raju, Bode, & Larsen, 1989), number of examinees (Rogers, 1990),

¹Paper presented at the meeting of AERA, Chicago, 1991.

²Laboratory of Psychometric and Evaluative Research Report No. 210.
Amherst, MA: University of Massachusetts, School of Education.

type of bias (i.e., uniform vs. non-uniform) (Rogers, 1990), level of homogeneity of the test (Clauser, Mazor, & Hambleton, 1990), etc.

Several of these studies have produced results which call into question the stability, if not the validity, of the Mantel-Haenszel results. For example, papers by Raju et al. (1989) and Clauser et al. (1990) report surprisingly high numbers of items identified as differentially functioning. This raises the question of whether these results may reflect false positive findings. Because these studies were based on actual test results, it is impossible to draw a final conclusion. However, in some preliminary follow-up research using simulated data, the MH statistic failed to identify a substantial number of the simulated biased items (Clauser & Mazor, 1990). Items were simulated using a three-parameter IRT model. The (item discrimination) a-parameters were equal for both reference and focal groups (randomly set from a pre-specified distribution). The (item difficulty) b-parameters (similarly chosen from an identified distribution) were set so that reference and focal groups differed by up to 0.6 standard deviations. Examination of the results suggested that the Mantel-Haenszel values increased as the difference in the b-parameter values increased, but that the size of this increase was clearly related to the item discrimination. Low discrimination items were associated with low Mantel-Haenszel values even when the difference in the b-parameters was large. High discrimination items produced higher Mantel-Haenszel values when compared across similarly divergent b-parameter values. Because the study was primarily intended to examine a different aspect of the Mantel-Haenszel statistic, the data did not allow a systematic examination of this value.

The present study extends this line of inquiry by examining effects of changes in the a- and b-parameters and accompanying interactions as well as the amount of simulated bias on the performance of the MH statistic. Previous studies (Hambleton & Rogers, 1989; Rogers, 1990) have shed significant light on the situation in which the a-parameters have unequal values for the two groups of interest resulting in item characteristic curves (ICCs) which cross. Because the MH statistic is a signed statistic, the positive and negative values created in such situations tend to cancel each other, making the MH procedure blind to non-uniform bias. By contrast, the present research is concerned with the conceptually simpler case where the a-values are equal for the reference and focal groups but high for some items and low for others. These differing a-values are examined in combination with both between group differences in the b-parameters and various overall levels of item difficulty.

In light of the considerable attention the Mantel-Haenszel statistic has received, surprisingly little research is available on this aspect of its performance. In the context of comparing the Mantel-Haenszel and logistic regression approaches to identification of differential item functioning, Rogers (1990) provided some insights into this aspect of the statistic's performance. She found significant results when varying both the item type and size of the simulated "bias." Item types included those with low and high a-parameters (i.e., .60 and 1.20) and high, medium, and low b-parameters (i.e., 1.50, 0.0, -1.50). The present study extends this research by examining a wider range of a- and b-parameter values and by examining them in the context of both equal and unequal ability distributions. The main question is: what are the statistical

characteristics of items which are known to be biased and which go undetected by the MH statistic?

Method

The current study uses simulated data produced using DATAGE (Hambleton & Rovinelli, 1973), a computer program to simulate examinee item response data fitting a one-dimensional logistic model. To produce the test into which the simulated "biased" items were placed, a- and b-parameters were taken from 59 randomly chosen items from the 1985 administration of the Graduate Management Admission Test (Kingston, Leary, & Wightman, 1988). These parameters were used to more closely approximate conditions found in practice. Sixteen items containing simulated bias were then added to the 59 to make a total test of 75 items. This test length was chosen because it is within the range occurring in typical standardized testing situations (e.g., achievement subtests) and yet long enough to reduce the instability which can be associated with Mantel-Haenszel results for shorter tests. It also allowed for a substantial number of items for study without making the percentage of biased items in the test greater than that which has been routinely identified in actual tests.

The c-parameters for all biased items were set at 0.20. The a-parameters for the studied items were set at 0.25, 0.60, 0.90, or 1.25. These values were crossed with four levels of simulated bias represented by differences in the b-parameters for focal and reference groups of 0.25, 0.50, 1.00, and 1.50. These 16 item parameter combinations (4 levels of item discrimination x 4 levels of b-value difference) were then staggered across five levels of item difficulty producing reference group b-parameter values of -2.50, -1.00, 0.00, 1.00, and 2.50. To allow for this number of

studied items (80), five simulated tests were required, each containing 16 biased items. Table 1 shows the a-parameter values, and the reference and focal group b-parameter values for each of the 80 studied items.

Figure 1 shows the reference and focal group item characteristic curves (ICCs) for four of these 80 items. ICC 1 and ICC 4 represent the extremes of simulated item bias in the study. ICC 2 and ICC 3 represent middle levels of simulated item bias. As the curves suggest, the level of simulated bias differs markedly across the studied items. The p value differences for the items (for the groups simulated with equal ability distributions) range from .01 to .52.

The examinee item responses for the 59 non-biased items were held constant for the five test simulations to prevent chance differences in these responses from influencing the effects under study. Responses were produced for 1,000 examinees in each group. Although this is a larger sample size than may be available to many researchers in practice, particularly during the screening stage for most tests, it was chosen because previous research has indicated that sample size is a significant variable with the Mantel-Haenszel statistic (Rogers, 1990). It was assumed that this sample size would increase the stability of the results and, by so doing, allow for a clearer focus on the variables of interest.

Ability distributions were created for the reference and focal groups so they would be equal and normally distributed with a mean of 0.0 and a standard deviation of 1.0. This arrangement is similar to that which is often encountered when using the MH statistic to assess for male-female differences. The simulations were then repeated using distributions that differed by 1.0 standard deviation. The distribution for the reference group remained as described above. The distribution for the focal group

had the same shape but with a lower mean. This arrangement was intended to simulate conditions sometimes found in item bias studies (see, for example, Hambleton & Rogers, 1989).

The MH procedure was then run on each of the ten sets of simulated test results.

Results

The specific items identified by the MH are indicated on Table 1. Table 2 summarizes results for the equal ability examinee groups. The procedure identified (at the .01 level of statistical significance) 49 of the 80 items which contained some level of simulated bias. One of the 59 "non-biased" items used in the five simulations was identified as differentially functioning. As the difference in the difficulty for focal and reference groups increased, the probability that the item would be identified as differentially functioning increased dramatically. A similar but less dramatic effect was noted for increases in the discrimination parameter. Although the absolute value of the difficulty parameter appeared to strongly influence the results, the impact was limited to the extreme upper range of the scale (i.e., b-values in the vicinity of +2.50).

Table 3 shows these results replicated for reference and focal groups with unequal ability distributions. Although the MH statistic identified fewer items as differentially functioning under this condition (46 out of 80), the pattern of results relative to the other variables was similar to those noted above. Table 4 shows ANOVA results for all of the studied items across both conditions of ability distribution.

Discussion

In general, these results appear to confirm the overall validity of the MH statistic to identify uniform differential item functioning. Approximately 60% of the "biased" items were identified. Only one "non-biased" item was identified. This is well within the range that might be predicted using a .01 significance level. As expected, the greater the difference in the difficulty parameters between groups, the more likely it was that an item would be identified. However, this pattern of performance of the MH statistic was highly influenced by the other variables examined. In general, an item with a given difference in difficulty was nearly 50% more likely to be identified as differentially functioning with a discrimination value of 0.90 or above as with 0.25. Although it seems necessary to include the discrimination value of an item in determining the impact of difficulty differences for the item, these results suggest that eliminating items during screening based on MH results is likely to differentially favor items of lower discrimination. In eliminating this percentage of the more highly discriminating items, a less biased test will be produced. Still, some bias may exist because it is contained in the less discriminating items where it remains undetected, even with large examinee samples.

Another variable of concern was that of the absolute value of the difficulty parameter of the studied item. This analysis examined the question of whether the MH statistic functioned equally effectively across the entire range of the difficulty scale. The clear answer is that it does not. The present results suggest that the procedure was essentially blind to differential functioning in the relatively most difficult items. Again, this result cannot be considered much of a surprise. Extremely difficult

items can only function differentially for the relatively few examinees at the upper end of the ability distributions. The majority of examinees in both groups will get the item wrong. Because the cells within the MH statistic are weighted by the number of examinees in each cell, differential functioning occurring within a few sparsely populated cells at the extreme end of the distribution will be missed. The question of whether ignoring differential functioning occurring under these conditions is appropriate depends on the use for which the test is intended. The weighting of the MH statistic allows identification of uniformly biased items at that part of the ability distribution where most of the examinees are functioning. Although this is generally advantageous, it could become problematic under several circumstances. First, if the ability distributions of examinees on which the test is screened are different than those on which the test will be employed; second, if it is important that the test be highly valid across the entire score range; and third, (and perhaps most importantly) in those conditions where a test is specifically intended to identify a relatively small number of highly competent examinees. Under this latter condition, it may be desirable to run the MH procedure on a selected group of scores from the most competent examinees. The loss of power associated with reducing the sample size is likely to be more than offset by focusing the analysis on the specific part of the ability distribution of interest.

The final variable considered was the relative level of ability for the focal and reference groups. Although the significant interactions make caution appropriate in interpreting the main effects of the ANOVA, there is additional evidence in the data to suggest that the disparity in ability between groups may have a meaningful impact on the MH results. A

comparison of the p-value differences between groups, associated with items identified or missed by the MH, suggests that the statistic may have greater sensitivity when used with groups of equal ability. The largest p-value difference associated with an item that was missed with the equal ability comparison was .08. This is in contrast to .15 for the unequal ability comparison. (P-values were estimated using equal ability distributions for groups of 2000 examinees with mean 0.0 and standard deviation of 1.0.)

Examination of Table 4 indicates that the main effects for each of the four variables considered are significant. In addition, four of the six two-way interactions are significant (at .01). Although this latter result may limit the interpretations placed on the main effects, it is consistent with our assumptions about the statistic. It is obvious that the main effects for a- and b-parameter values are partially dependent on the difference in the b-parameters between groups. If this difference equals 0.00, it is unlikely that any level of difficulty or discrimination will produce a significant MH value. Similarly, if the significance of the b-parameter value is explained by the fact that potential bias in very difficult items cannot be identified because there are too few examinee functioning in that part of the ability distribution, then shifting the ability distribution (to decrease the ability for the focal group) should be expected to exaggerate this effect. An examination of the data supports this explanation. The five items identified on the equal ability distribution comparison and missed on the unequal were all moderate to very difficult (i.e., four reference group b-values of 1.00 and one of 2.50). The two items identified on the unequal ability distribution comparison and

missed on the equal were extremely easy (i.e., both reference group b-values were -2.50).

Overall, this research supports the continued use of the Mantel-Haenszel statistic for identifying differential item functioning. It does, however, suggest that caution should be used in attempting to identify such "uniform bias" in extremely difficult items. This problem may be exaggerated in cases where the ability distributions for the focal and reference groups are substantially different. The possible problems referred to above regarding the impact of the a-parameter may merit further research.

References

- Clauser, B. E., & Mazor, K. (1990, May). The effects of score group size and ability distribution on Mantel-Haenszel results. Paper presented at the meeting of NEERO, Rockport, Maine.
- Clauser, B. E., Mazor, K., & Hambleton, R. K. (1990). The influence of test homogeneity on the identification of differentially functioning test items using the Mantel-Haenszel statistic (Laboratory of Psychometric and Evaluative Research Report No. 198). Amherst, MA: University of Massachusetts, School of Education.
- Hambleton, R. K., & Rogers, H. J. (1989). Detecting potentially biased test items: Comparison of IRT area and Mantel-Haenszel methods. Applied Measurement in Education, 2(4), 313-334.
- Hambleton, R. K., & Rovinelli, R. J. (1973). A Fortran IV program for generating examinee response data from logistic test models. Behavior Science, 17, 73-74.
- Holland, P., & Thayer, D. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), Test validity. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Kingston, N., Leary, L., & Wightman, L. (1988). An exploratory study of the applicability of item response theory methods to the Graduate Management Admission Test (GMAC Occasional Papers). Princeton, NJ: Graduate Management Admission Council.
- Raju, N. S., Bode, R. K., & Larsen, V. S. (1989). An empirical assessment of the Mantel-Haenszel statistic for studying differential item performance. Applied Measurement in Education, 2(1), 1-13.
- Rogers, H. J. Item bias investigations with logistic regression. Unpublished doctoral dissertation, University of Massachusetts, 1990.

Table 1

Parameter Values for the 80 Items Containing Simulated Bias

Item	a	Test 1		Test 2		Test 3		Test 4		Test 5	
		b _R	b _F	b _R	b _F	b _R	b _F	b _R	b _F	b _R	b _F
1	.25	-2.50	-2.25	2.50	2.75	1.00	1.25	0.00	0.25	-1.00	-0.75
2	.25	-2.50	-2.00	2.50	3.00	1.00	1.50	0.00	0.50	-1.00	-0.50
3	.25	-2.50	-1.50♦*	2.50	3.50	1.00	2.00 *	0.00	1.00♦*	-1.00	0.00♦*
4	.25	-2.50	-1.00♦*	2.50	4.00♦*	1.00	2.50♦*	0.00	1.50♦*	-1.00	0.50♦*
5	.60	-1.00	-0.75	-2.50	-2.25♦	2.50	2.75	1.00	1.25 *	0.00	0.25
6	.60	-1.00	-0.50♦*	-2.50	-2.00♦*	2.50	3.00	1.00	1.50	0.00	0.50♦*
7	.60	-1.00	0.00♦*	-2.50	-1.50♦*	2.50	3.50 *	1.00	2.00♦*	0.00	1.00♦*
8	.60	-1.00	0.50♦*	-2.50	-1.00♦*	2.50	4.00	1.00	2.50♦*	0.00	1.50♦*
9	.90	0.00	0.25♦*	-1.00	-0.75♦*	-2.50	-2.25	2.50	2.75	1.00	1.25♦*
10	.90	0.00	0.50♦*	-1.00	-0.50♦*	-2.50	-2.00♦*	2.50	3.00	1.00	1.50♦*
11	.90	0.00	1.00♦*	-1.00	0.00♦*	-2.50	-1.50♦*	2.50	3.50	1.00	2.00 *
12	.90	0.00	1.50♦*	-1.00	0.50♦*	-2.50	-1.00♦*	2.50	4.00	1.00	2.50 *
13	1.25	1.00	1.25	0.00	0.25	-1.00	-0.75♦*	-2.50	-2.25♦	2.50	2.75
14	1.25	1.00	1.50	0.00	0.50♦*	-1.00	-0.50♦*	-2.50	-2.00♦*	2.50	3.00
15	1.25	1.00	2.00♦*	0.00	1.00♦*	-1.00	0.00♦*	-2.50	-1.50♦*	2.50	3.50
16	1.25	1.00	2.50♦*	0.00	1.50♦*	-1.00	0.50♦*	-2.50	-1.00♦*	2.50	4.00

*Items identified by the MH on the equal ability distribution comparison

♦Items identified by the MH on the unequal ability distribution comparison

Table 2

Mantel-Haenszel Results for Reference and Focal Groups with Equal Ability Distributions

Condition	Items Identified	
	Number	Percent
b-Parameter Difference		
.25	5	25
.50	10	50
1.00	17	85
1.50	17	85
a-Parameter Value		
.25	9	45
.60	13	65
.90	15	75
1.25	12	60
b-Parameter Value (Reference Group)		
-2.50	11	69
-1.00	13	81
0.00	12	75
1.00	11	69
2.50	2	13

Table 3

Mantel-Haenszel Results for Reference and Focal Groups with Unequal Ability Distributions

Condition	Items Identified	
	Number	Percent
b-Parameter Difference		
.25	6	30
.50	10	50
1.00	14	70
1.50	16	80
a-Parameter Value		
.25	8	40
.60	12	60
.90	13	65
1.25	13	65
b-Parameter Value (Reference Group)		
-2.50	13	81
-1.00	13	81
0.00	12	75
1.00	7	44
2.50	1	6

Table 4

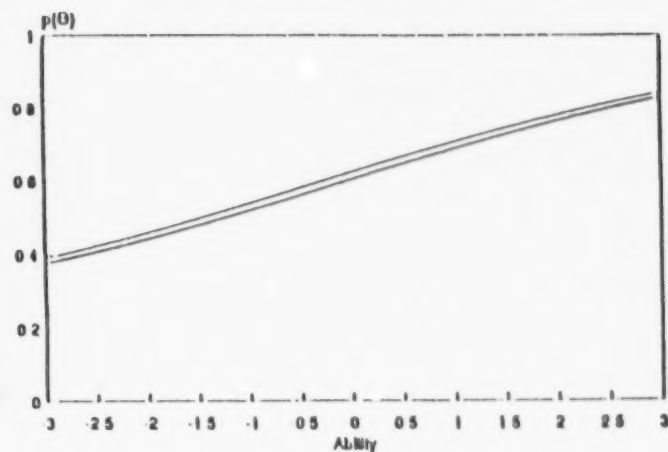
Analysis of Variance of Mantel-Haenszel Statistics

Factor	F	P
b-Parameter Difference	94.924	.000*
a-Parameter Value	29.441	.000*
b-Parameter Value	39.673	.000*
Ability Distribution	6.516	.012*
b-Par. Diff. x a-Par. Val.	6.964	.000*
b-Par. Diff. x b-Par. Val.	11.722	.000*
b-Par. Diff. x Ability Dist.	2.827	.042
a-Par. Val. x b-Par. Val.	7.022	.000*
a-Par. Val. x Ability Dist.	.242	.867
b-Par. Val. x Ability Dist.	3.577	.009*

* P < .01

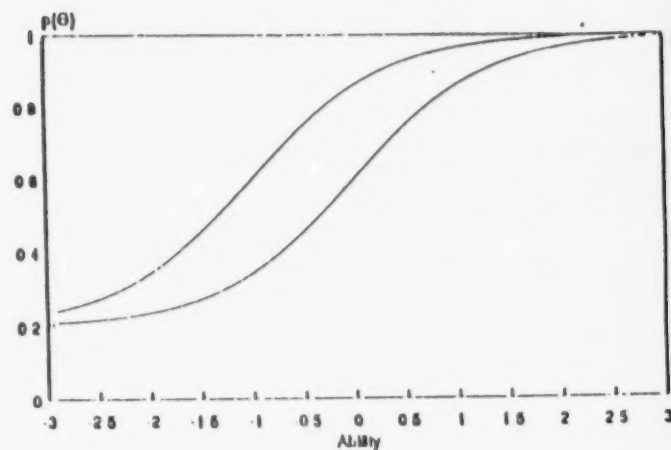
ICC 1

$a = .25$, $b \text{ difference} = .25$



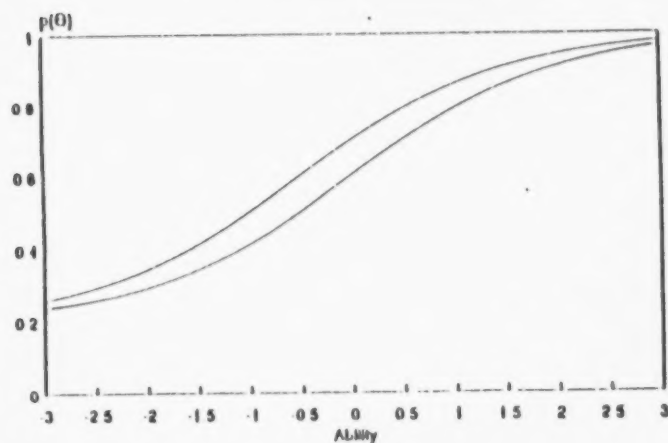
ICC 3

$a = .90$, $b \text{ difference} = 1.00$



ICC 2

$a = .60$, $b \text{ difference} = .50$



ICC 4

$a = 1.25$, $b \text{ difference} = 1.50$

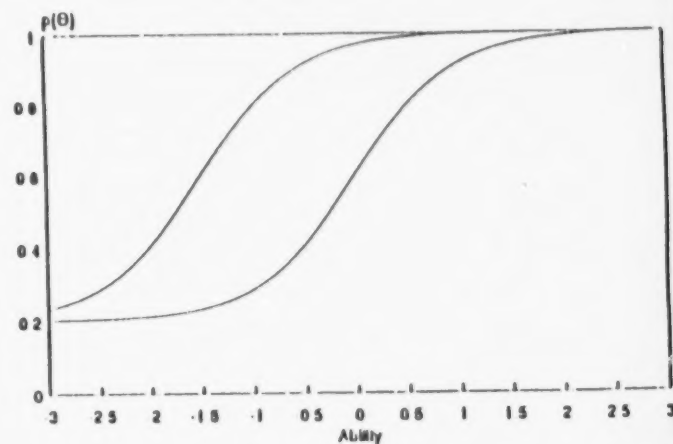


Figure 1. A sample of the biased items in the computer simulation study.